

**Slide 1 of 24**

Title Slide: Data Collected through the Postsecondary Studies

**Slide 2 of 24**

This module focuses on the three types of source data that are collected through the postsecondary education studies and how those source data are processed to create the derived files that are used by QuickStats and PowerStats and researchers with access to restricted-use micro-level data. It includes data collected from postsecondary institutions, student interview data, and data gathered from other administrative sources.

**Slide 3 of 24**

As noted earlier, NCES's postsecondary studies draw from three broad classes of data.

The first data collected from institutions, include: (a) institutional characteristics from the Integrated Postsecondary Education Data System, or IPEDS, (b) data from campus student information systems, including academic and financial aid data, and (c) for longitudinal studies, student transcripts.

The second major data source for NPSAS, BPS, and B&B is the student interview. As a reminder, these web-based interviews are either self-administered or completed with the assistance of a trained telephone interviewer.

Finally, NCES postsecondary studies gather administrative records about sampled students from both internal and external sources, including the Department's Central Processing System, which houses FAFSA data, and the National Student Loan Data System, which tracks student loan and Pell grant disbursements; as well as data from the National Student Clearinghouse and admission testing firms.

**Slide 4 of 24**

In NPSAS years, data are collected from institutions in two phases.

First, sampled institutions are asked to provide lists of all students enrolled during the study year. Those lists include all information needed for NPSAS sampling and sampling for that cycle's longitudinal spin-off, either BPS or B&B. These indicators may include field of study, academic class level, indication of "first-time beginning" or "expected graduate" status, and other key student characteristics.

Using enrollment lists, specific students are sampled. NCES then recontacts participating institutions, asking for detailed information about students who have been sampled. This includes contacting information, detailed enrollment and academic histories, and financial aid information.

## **Data Collected Through the Postsecondary Studies**

Additional institutional data are collected related to the students selected to participate in longitudinal studies, including their transcripts. For the BPS cohort starting in 2012, NCES intends to gather student record data from institutions for each year of a student's enrollment.

Because institutional participation is critical to the success of all NCES postsecondary studies, NCES and its data collection contractors work closely with sampled institutions. This includes technical assistance with Institutional Review Board compliance, data extraction and collection, and any other issue that might hinder timely cooperation.

### **Slide 5 of 24**

NCES provides institutions a secure, web-based application to provide student records data about each sampled student. An example is shown here. Rather than manually entering data, institutions may also upload CSV files containing student records data to a secure NCES server.

### **Slide 6 of 24**

The second major data source used in postsecondary studies is the student interview.

The general process of student interview data collection includes:

- (1) Verifying the quality of respondent contacting information, including matching to publicly available and commercial databases;
- (2) In the early phase of data collection, students are contacted via postal and electronic mail to encourage response to the student interview via the Web. Some students may also receive prompting telephone calls in this early phase to help ensure their response;
- (3) In later phases of data collection, trained interviewers begin to call any cellular or landline telephone number on file for sampled students in an effort to encourage Web response or to complete the interview on-the-spot;
- (4) In the final phase of data collection, some studies offer remaining nonrespondents an abbreviated interview, which is shorter in length and focuses on a subset of key data elements.

In recognition of the value of respondents' time, NCES offers a monetary token of completion at the conclusion of the student interview.

The specific contacting procedures used for a given study are designed to yield high response rates overall, but most particularly among students who, if they failed to respond to the study, would likely alter, or bias, study results. Reduction of this nonresponse bias is a key component of all NCES data collection efforts.

### Slide 7 of 24

This screen provides an example of a web-based student interview question. In this example, a student sampled to participate in B&B has visited the secure B&B data collection website, logged in with their unique credentials, and is now in the midst of the student questionnaire.

### Slide 8 of 24

The final category of source data includes records pulled from other administrative data sources. Here are examples of these administrative sources to which postsecondary study sample members are matched. For example, NCES works with College Board to obtain SAT test scores for both BPS and B&B sample members. These scores will appear in the final data files for you to use in your analyses. For an exhaustive list of administrative data sources for each study, please refer to the studies' methodology reports; which can be accessed by clicking on the underlined screen text, or from the summary and resources screen at the end of this module.

### Slide 9 of 24

Although many NCES studies will use questionnaire respondents as their unit of analysis, postsecondary studies are unique in that a data-driven, rather than questionnaire-driven, definition of response is used. Because key data elements for each study are frequently available from multiple sources, including the interview, the unit of analysis for NPSAS, BPS, and B&B is a **study member**. Students are defined as study members when a subset of key characteristics about a student can be derived from any source. These characteristics include the student type, birth date or age, gender, and at least eight of the following 15 variables: dependency status, marital status, any dependents, income, expected family contribution (EFC), degree program, class level, first-time beginner status, months enrolled, tuition, received financial aid, received non-federal aid, student budget, race, and parent education.

### Slide 10 of 24

We now can examine the data collected from postsecondary institutions, student interviews, and other administrative data sources in terms of the specific postsecondary studies that they inform. The NPSAS, BPS, and B&B all obtain data from each of the three major data sources. For example, BPS institution-level source data come from IPEDS, student records, and transcripts. BPS also obtains source data from a student interview data collection (conducted first through NPSAS and then during a 3rd and 6th year follow-up). Lastly, BPS obtains source data from existing administrative databases including the Central Processing System – which contains FAFSA information; NSLDS; National Student Clearinghouse enrollment information; and ACT and SAT college admissions test scores.

This table shows the specific source files across the three major postsecondary education data collection efforts that may be available as source files on your restricted-use micro-level data file. In the next portion of this module we will discuss how NCES

uses these sources of data to derive analysis variables available in QuickStats, PowerStats, and on the restricted-use CD.

### Slide 11 of 24

NCES has developed procedures for processing the source data provided by institutions, students, and administrative records and aggregating them into files created specifically for analysis purposes. These derived files have been edited for errors and inconsistencies, and missingness has been almost eliminated using imputation from other available information from comparable sources. For this reason, and many more, we suggest that restricted-use data licensees use variables from the derived file whenever possible. PowerStats reads from the derived file, which can be found at the top-level of the restricted-use file CDs. In addition to the derived file, restricted-use data CDs contain the 'source' variable files from which the derived variables are derived. Again, we advise that licensees use source variables only when the derived file doesn't contain what is needed.

In the screens that follow, we will work through an example of how data from a variety of source files come together to create a derived variable that allows researchers to understand a study member's aid application status.

### Slide 12 of 24

Here is an example of how a composite or, as we call it, a "derived variable" is created. On this page we have provided you with four source variables from four separate postsecondary education data systems. The first variable, 'INCPS' spelled I-N-C-P-S, is an administrative variable from the Central Processing System, which is the system that stores the FAFSA application data. The second variable, 'Did You Apply', is a student interview variable supplied directly by a study member relating to whether or not they applied for student financial aid. The third, 'Total Federal Aid', is another administrative variable from the National Student Loan Data System, which stores data relating to federal student loan disbursements and repayment. The fourth variable, 'Total Aid (All Sources)', is a variable provided by the student's institution as part of the uploaded student records collection.

If you opened any of the source files from these data sources on your restricted-use micro-level data files, these are just a few of the variables that you would see. As you can see, there are a lot of missing, or nonresponse, or no values in the individual variables of these source file data sets. It becomes necessary to bring all the relevant variables about student aid together to understand it in a more robust way than any of the individual source files describes student aid. We have done this for you through creating the derived variables, so we recommend using these before turning to the "source" files.

NCES takes this into account when we derive variables, we use multiple sources of relevant data to derive a response to the question of interest, which in this case is, "Did a study member apply for either Federal or any student aid?". Accordingly, in this

example we show you how the derived variables FEDAPP, spelled F-E-D-A-P-P, (apply for federal aid) and AIDAPP, spelled A-I-D-A-P-P (apply for any aid) are created to address the 'Aid' question. All the subsequent slides presented in this section of the module, highlight our approach to deriving variables for analysis. That is, what we look at and how our consideration of data provided by multiple sources leads to a response to a question – and a derived variable.

So let's begin. An examination of the 'Applies to' field for INCPS, the first variable we will consider as we derive our two aid application status variables, is a flag that will inform us which study members have information contained within the CPS. This means that if a student applied for federal student aid, INCPS is set to 'yes' or '1'. If CPS didn't have a record of them applying for aid, the field is set to 'missing'.

### Slide 13 of 24

If the student applied for Federal Aid, a record is created in the Department of Education's Central Processing System. If a record is present, FEDAPP is set to "Yes."

As we have seen in this example, there are three study members who have a record in the CPS source file, so we will indicate their 'FEDAPP' derived variable status as 'Yes'.

As you can see here the red '1s' in the INCPS column now correspond with a 'Yes' in the FEDAPP derived variable column.

### Slide 14 of 24

Next, we will consider the total federal aid column in this table. If the student received any federal aid or, in other words, Total Federal Aid is greater than \$0 according to the National Student Loan Data System, FEDAPP is set to "Yes" even if no CPS record is present.

There are now two more study members who are now marked as 'Yes' in the FEDAPP derived variable column.

### Slide 15 of 24

Next, for any of the remaining study members who stated that they didn't apply for financial aid—as noted here with a 'no' response—FEDAPP is set to "No". It is important to note that even if a student replied 'no' to the "did you apply" question, but is found to have a record in CPS (which is not shown here) or has in fact received federal financial aid according to other administrative records, then "Applied for federal aid" will be set to 'Yes'. Also, if the respondent said that they did apply for student aid, but both INCPS and 'Total Federal Aid' are both missing, it is assumed that they must have applied for non-federal student aid, so FEDAPP is set to "No."

**Slide 16 of 24**

Now, we can move to the second derived variable, Applied for Any Aid or 'AIDAPP'. If a student had applied for federal aid or FEDAPP is "Yes", then 'applied for any aid' or AIDAPP is logically set to "Yes".

As you can see, all the 'Yes's' that we just derived in creating 'FEDAPP' are used to begin deriving this new variable, called AIDAPP, for analysis.

**Slide 17 of 24**

Next, if a student received any aid whatsoever - that is any time that Total Aid > \$0 - as reported by their postsecondary institutions, then AIDAPP is set to "Yes" if it hasn't already been done so.

Presuming the student applied for aid through non-federal entities, the number of study members who are considered 'Yes's' for AIDAPP should be greater than the number who have 'Yes's' for FEDAPP. In other words, AIDAPP "yes's" include all those who applied for federal aid plus those who applied for non-federal aid.

**Slide 18 of 24**

Next, if a student indicated in the interview that they applied for aid, but INCPS, Total Federal Aid, and Total Aid are all set to missing, then it is assumed that they applied for some form of non-federal aid and AIDAPP is set to "Yes".

As you can see on your screen, the Yes in the interview variable 'Did you Apply' in the second column is highlighted, as well as the corresponding 'Yes' in the derived AIDAPP variable column.

**Slide 19 of 24**

Next, if a student reported that they didn't apply for student aid and there is no other indication that they had, then AIDAPP is set to "No".

**Slide 20 of 24**

The last step of the derivation process involves only cases that have missing data across all of the relevant source files needed to create the derived variable of interest. In this step, cases that have missing data are statistically imputed.

This example showed you an abbreviated logical process that NCES uses to create and derive variables for analysis. As you can see, there are many variables across the postsecondary education data collections and data sources that may appear similar but are actually very different. It is imperative to always consult the README file and the codebook associated with your derived file of interest to ensure proper analysis of the postsecondary education data.

**Slide 21 of 24**

The imputation procedures involves a four-step process: In the first step, missing variables are logically imputed as you have seen in the previous example; that is, students who did not respond to the question ‘Did you apply for aid?’ but received aid were assumed to have applied.

In the second step, the criteria used to match variables into imputation classes to stratify the dataset are identified so that all imputations are processed independently within each class of similar respondents.

In the third step, the weighted sequential hot deck process is implemented, whereby missing data were replaced with valid data from donor records that most closely match the recipients with respect to the matching criteria.

The term hot deck refers to the fact that the set of potential donors changes for each recipient. In contrast, cold deck imputation defines one static set of donors for all recipients. In all such imputation schemes, the selection of the donor from the entire deck is a random process.

In the fourth step, a cyclic n-partition hot deck process is implemented to iteratively cycle through a set number of partitioned hot decks. See “Large-Scale Imputation for Complex Surveys” by Marker, Judkins, and Winglee (2002) for more information which can be accessed by clicking on the underlined screen text.

**Slide 22 of 24**

This is a screenshot from PowerStats. PowerStats is another resource where postsecondary education restricted-use micro-level data users can find information about derived variables. This same information is contained within the codebooks and the restricted use software provided with your restricted-use CD.

The programming notes that are provided within the data file documentation show how the derived variable was created. Sometimes you might see SAS code, or other syntax, detailing the derivation of the variable of interest. Which data sources were used in the creation of each derived variable are also included. For more information about PowerStats, click on the corresponding underlined screen text.

**Slide 23 of 24**

This module has described the three types of source data that are collected through the postsecondary education studies. Additionally, this module has described how those source data are processed to create the derived files that are used by QuickStats, PowerStats, and researchers with access to restricted-use micro-level data.

**Slide 24 of 24**

This module has also provided resources that can be accessed through the DLDT system and/or on the NCES website.

You may now proceed to the next module in the series, or click the exit button to return to the landing page.